Journal of Computer Science

Research Article

Comparative Study of BERT Based Architectures for Multi Task News Classification and Threat Detection

A. Mustain Billah and Sani M Isa

Department of Computer Science, Bina Nusantara Graduate Program Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia

Article history
Received: 02-06-2025
Revised: 15-09-2025
Accepted: 07-10-2025

Corresponding Author:
A. Mustain Billah
Department of Computer Science, Bina Nusantara Graduate Program-Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia
Email: a.billah@binus.ac.id

Abstract: We present a comparative analysis of single-task and multi-task BERT-based models for Indonesian news classification across two objectives: category prediction (five classes) and threat detection (binary). Using 8,951 annotated news titles, single-task baselines achieved weighted F1 scores of 0.84±0.01 (category) and 0.87±0.00 (threat). Multi-task hybrids integrating CNN, LSTM, and Bi-LSTM layers performed comparably overall and improved minority class threat detection, with the BERT CNN variant attaining the highest threat F1 (0.88±0.01). Per-class results confirmed that the Ideology category, represented by only 279 samples, remained the most challenging. Efficiency benchmarks on an NVIDIA L4 demonstrated practical feasibility, with batch size 32 throughput of approximately 450-470 items per second (equivalent to 2.1 2.2 ms per title) and single-item latency of around 67-69 ms. Training times ranged from 138±12 to 193±22 s across seeds. These findings indicate that multi-task BERT hybrids can improve threat detection while sustaining near real-time throughput, supporting their applicability in large scale monitoring of Indonesian news streams.

Keywords: Multi-Task Learning, BERT, CNN, LSTM, Bi-LSTM, News Classification, Comparative Study

Introduction

Automatic classification of news articles plays a crucial role in enabling timely information retrieval and threat assessment in rapidly evolving media environments. Transformer-based models such as BERT (Devlin et al., 2019) have set new benchmarks in single-task text classification by capturing rich, contextualized embeddings. However, most studies focus on either topic categorization or threat detection in isolation. In real-world scenarios, such as monitoring online news for ideological bias and security threats, systems must perform multiple classification tasks simultaneously, placing heavy demands on both accuracy and computational efficiency.

Multi-task learning serves as an effective strategy by promoting representation sharing among interconnected tasks, which can improve generalization and reduce redundant computation (Crawshaw, 2020). Complementary work has sought to enhance BERT's embeddings with convolutional layers for local n-gram feature extraction (Abas et al., 2022) or with recurrent modules such as LSTM and Bi-LSTM to capture sequential dependencies (Belaroussi et al., 2025). Despite these promising hybrid approaches, there remains a lack of systematic comparison among them under a unified multi-task framework.

To fill this gap, we conduct a comprehensive comparative study of four transformer-based architectures:

1. Single-task BERT, trained separately for category classification and threat detection
2. BERT-CNN, integrating convolutional filters for local feature extraction
3. BERT-LSTM, appending LSTM layers to model long-term dependencies
4. BERT-Bi-LSTM, employing bidirectional LSTM for richer context

The main contributions of this work are:

1. A systematic comparison of single-task and multi task BERT-based models, including hybrid CNN and LSTM variants, for simultaneous news category and threat classification

2. Empirical evaluation on a large-scale Indonesian news dataset, demonstrating that multi-task learning can achieve competitive performance while reducing inference time
3. An operational efficiency analysis, explicitly reporting both throughput-normalized latency with batch size 32 and single-item latency with batch size 1, confirming near-real-time feasibility for deployment

Related Work

Early efforts to augment BERT's embeddings with convolutional or recurrent encoders have shown clear benefits for single tasks. Abas et al. (2022) demonstrated that stacking convolutional filters on top of BERT can capture local n-gram patterns for emotion detection. Keya et al. (2023) proposed the Tri BERT CNN-LSTM model for fake-news detection and achieved state-of-the-art accuracy on English benchmarks. Mehta et al. (2022) applied a BERT CNN architecture to hate-speech detection and reported an F1 increase from 0.85 to 0.95. Rai et al. (2022) observed modest gains of 1 to 2% when appending a unidirectional LSTM layer to BERT for sentiment classification.

Shah et al. (2024) introduced a multi-task BERT framework with two task-specific heads to predict news category and sentiment concurrently, achieving 98% topic classification accuracy and 94% sentiment accuracy on a dataset of 3,263 news articles. Follow-up work extended MTL to sarcasm + sentiment (El Mahdaouy et al., 2021), peer-feedback classification (Jia et al., 2021), and hierarchical offensive-language detection (Dai et al., 2020). Collectively, these papers confirm that task-level supervision can be consolidated without sacrificing performance, provided loss balancing and task similarity are carefully managed.

Most Indonesian work to date tackles one textual objective at a time. Hendrawan et al. (2020) combined BERT and Bi-LSTM for multi-label hate-speech detection on Twitter. Lin and Nuha (2023) embedded BERT and TCN for sentiment analysis (85.13% accuracy). Indo BERT + Bi-LSTM reduced COVID-19 misinformation with 87.02% accuracy (Faisal and Mahendra, 2022). A 2024 transfer-learning study on Indonesian fake-news classification likewise remained single-task (Praha et al., 2024).

While prior multi-task BERT studies have typically paired topic classification with sentiment, sarcasm, or offensive-language detection (Dai et al., 2020; El Mahdaouy et al., 2021; Shah et al., 2024), our study makes four distinct contributions. First, it focuses on Indonesian-language data, a low-resource context where dual-labeled corpora remains scarce. Second, classification is performed solely on short news titles rather than full articles, making efficiency and latency especially critical. Third, the tasks are defined by the Indonesian IPOLEKSOSBUDHANKAM taxonomy in combination with binary threat detection, a pairing not previously explored in prior work. Finally, we provide a detailed operational evaluation of inference, explicitly reporting both throughput-normalized latency and single-item latency to assess real-time viability. Together, these design choices position our contribution as a unique comparative and efficiency-focused evaluation of multi-task BERT hybrids in Indonesian news.

Methods

Multi-Task Learning

Multi-task classification leverages parameter sharing between related tasks. A pre-trained BERT model is taken and used as the base layer or backbone. Additional layers are then added for each different task. Thus, the model can learn a general representation of the data through the base layer and task-specific representations through the additional layers (Ruder, 2017).

In this study, the two tasks carried out by the model include news category classification and threat element detection, both performed at the document level. For the plain BERT-MT we use the [CLS] embedding and for hybrid variants (BERT-CNN/LSTM/Bi-LSTM) we use the full token sequence (last hidden state) before the task heads. Each Dense layer produces predictions for each task, namely news category classification and determining whether the news contains threat elements.

The loss functions of each task are combined into one overall loss function which is used to optimize the model. In this way, the model is optimized using information from every task concurrently during training. The gradients of this joint loss function are used to update the weights of the BERT model and additional layers simultaneously (Zhang and Yang, 2021):

$$L_{MTL} = \gamma_1 . L_1 + \gamma_2 . L_2 \quad (1)$$

Where:

- L_1 as the loss value of news category classification
- L_2 as the loss value of threat classification
- γ_1 as the weight value for news classification task
- γ_2 as the weight value for threat classification task

Multi-Task Variants

A single Indo BERT encoder first generates contextual embeddings, which are then processed by one of three interchangeable hybrid layers: A convolutional module, a unidirectional LSTM, or a bidirectional LSTM. This design enables a controlled comparison of local n-gram feature extraction, sequential modelling, and bidirectional context representation. The resulting feature vectors are passed to two task-specific classification heads, which

consist of a five-way softmax layer for topic categorization and two-logit softmax for threat detection.

The losses from each head are combined into a joint multi-task objective, L_{MTL} that is back-propagated through both the task heads and the shared encoder (Fig. 1 and Table 1).

The training procedure begins by initializing the selected model variant, the joint task weights (γ_1, γ_2), the Adam W optimizer with its learning rate schedule, and the early-stopping parameters. During each epoch, the dataset is divided into mini-batches. For each batch, raw text is tokenized and converted into BERT input tensors consisting of token IDs and attention masks. These tensors are then fed into the shared Indo BERT encoder to produce contextual embeddings. Depending on the chosen variant, the embeddings pass through either the identity mapping, a one-dimensional convolutional block, a unidirectional LSTM, or a bidirectional LSTM.

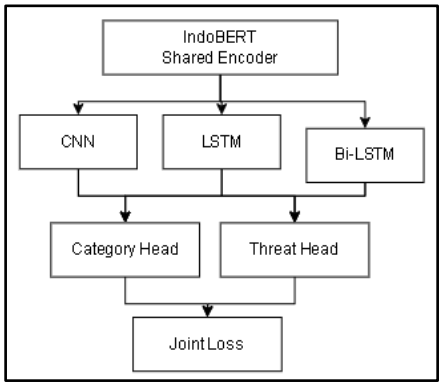

Fig. 1: Proposed multi-task hybrid architecture

Table 1: Pseudocode for multi-task hybrid BERT training

Step	Operation
Inputs	• Model variant $M \in$ {BERT-Threat, BERT-Category, BERT-MT, BERT-CNN, BERT-LSTM, BERT-Bi-LSTM} • Training data $D = \left\{\left(x_i, y_i^{(c)}\ y_i^{(t)}\right)\right\}$, where $y^{(c)}$ is category label and $y^{(t)}$ is threat label • Loss weight y_1 , y_2 • Optimizer Opt = AdamW • Learning rate η • Number of epochs E
1	Forward Pass: • tokens, mask ← tokenize(x) • H ←BERT(tokens,mask) • $F \leftarrow HybridLayer_M(H)$ • $p^{(c)}, p^{(t)} \leftarrow TaskHeads(F)$
2	Compute losses: $L_c sparse\ categorical\ crossentropy\ (p^{(c)}, B.y^{(c)})$ $L_t sparse\ categorical\ crossentropy\ (p^{(t)}, B.y^{(t)})$ $L = y_1.L_c + y_2.L_t$
3	Backward & update: with GradientTape(): preds ← M(batch) L ← compute_loss(y, preds) grads ← ∇L wrt M.parameters Opt.apply_gradients(grads, M.parameters)
4	Early stopping: If early_stopping_criteria_met() then break

The resulting feature vector is directed to two separate classification heads: A softmax head for topic categorization (five classes) and a softmax head for threat detection (two classes).

Once the outputs are produced, sparse categorical cross-entropy is applied separately to the topic classifier and the threat classifier, and the resulting losses are combined into a single multi-task objective, as presented in Formula (1). In our experiments, we set $\gamma_1 = 1.0$ and $\gamma_2 = 1.0$, thereby giving equal importance to both tasks. Equal weighting was adopted as a baseline. This choice also ensured a fair comparison across variants, as preliminary trials with skewed weighting did not yield measurable improvements. Gradients of the joint loss are back-propagated through both the shared encoder and the task heads, and the optimizer updates all parameters accordingly.

The Adam W optimizer was selected because it decouples weight decay from gradient updates and has

been demonstrated to improve stability and effectiveness in transformer fine-tuning (Loshchilov and Hutter, 2019). To balance efficiency and prevent overfitting, training was halted if no improvement was observed for two consecutive epochs, in line with best practices for transformer-based classification (Devlin et al., 2019). By applying identical optimization and stopping criteria across all architectures, we ensured that the comparative evaluation of CNN, LSTM, and Bi-LSTM hybrids remained unbiased and reproducible. While this study adopted equal weighting for clarity and comparability, future work may explore adaptive task-weighting strategies such as uncertainty weighting or Grad Norm to further optimize performance.

BERT-CNN

Figure 2 shows the multi-task architecture that uses BERT as a shared encoder to produce contextual embeddings for each input token. The last hidden state output from BERT is passed to a Conv1D layer with 250 filters (kernel size = 3) and a Re LU activation. This convolutional layer is included to capture local n-gram patterns and to emphasize phrase level features that complement the global context produced by self-attention. After convolution, a GlobalMaxPooling1D layer is applied to reduce each feature map to a fixed-length 250 dimensional vector. This pooled vector is then routed into two classification pipelines. In the first pipeline, two dense layers (category dense then category output) predict one of five topic labels. In the second pipeline, threat dense and threat output produce a two-class threat distribution. Parameter sharing is leveraged in both the BERT encoder and the convolutional block while task-specific representations are learned by each classification head.

During hyper-parameter tuning, the number of convolutional filters was varied among {300, 250, 200} to investigate trade-offs between representational capacity and computational efficiency. After evaluation, the configuration with 250 filters (kernel size = 3) was selected as the final setting, since it provided the best balance of accuracy and computational cost. This decision aligns with prior findings that mid-range filter sizes are effective for capturing local n-gram patterns without excessive complexity (Murfi et al., 2024). Accordingly, Table 2 reports the explored search space, while Figure 2 depicts the final chosen configuration with 250 filters.

BERT-LSTM

Figure 3 shows the multi-task architecture that uses BERT as a shared encoder to generate a sequence of 768-dimensional embeddings in last hidden state. Token indices and attention masks are fed into the shared encoder, and the resulting embeddings are processed by a single-layer unidirectional LSTM with 300 hidden units, yielding a sequence of 300-dimensional vectors. The final time-step output is extracted as a 300-dimensional feature vector.

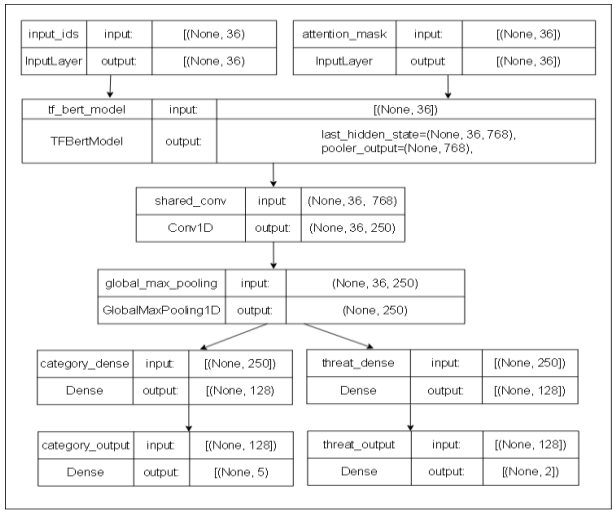

Fig. 2: Proposed BERT-CNN architecture

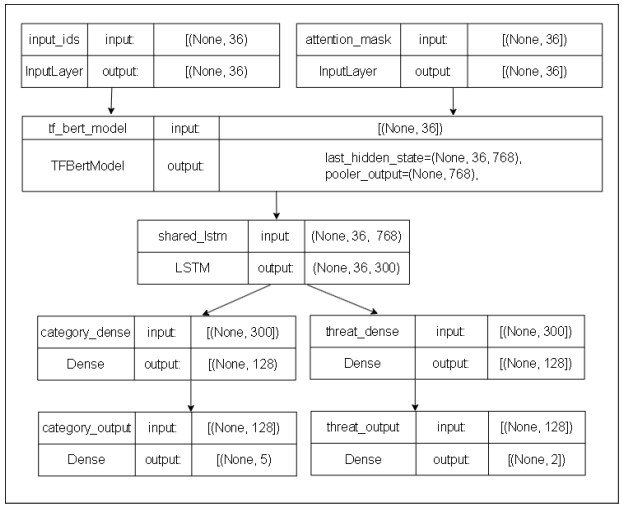

Fig. 3: Proposed BERT-LSTM architecture

This feature vector is then routed into two task heads: A 128-unit dense layer with softmax activation for topic categorization (five classes) and another 128-unit dense layer with softmax activation (two logits) for threat detection. During training, a joint multi-task loss is back-propagated through both the shared encoder and the classification heads.

BERT-Bi-LSTM

Token indices and attention masks are first fed into the shared Indo BERT encoder to produce a sequence of 768-dimensional embeddings (last hidden state). These embeddings are then processed by a bidirectional LSTM with 300 units in each direction, yielding a sequence of

DOI: 10.3844/jcssp.2026.1073.1082

600-dimensional vectors. The vector at the final time step is extracted as a 600-dimensional feature representation.

This feature is routed into two task-specific heads. For topic categorization, a 128-unit dense layer followed by a softmax activation predicts one of five news categories. For threat detection, a parallel 128-unit dense layer with a softmax activation over two logits estimates the probability of threat versus non-threat. A joint multi-task loss is then back-propagated through both heads and the shared encoder during training (Fig. 4)

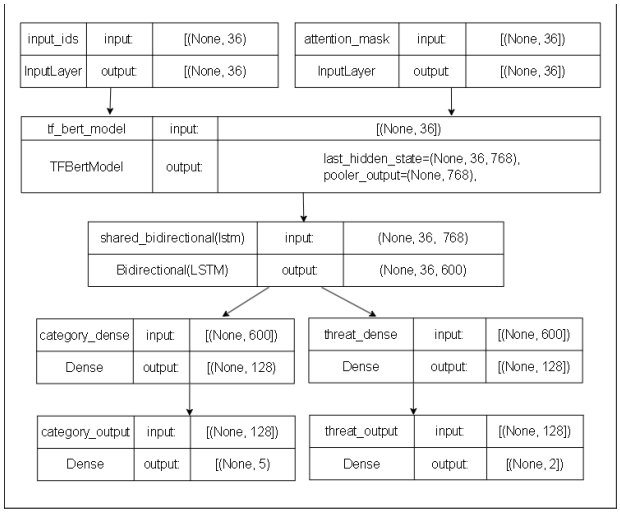

Fig. 4: Proposed BERT-Bi-LSTM architecture

Experiment Setup

Dataset Collection

We compiled a list of 500 Indonesian news RSS feeds drawn from public web searches and the Indonesian Press Council's media listings. Headlines were scraped hourly for one week, yielding 8,951 titles. Each title is associated with its outlet and timestamp of first appearance. (Figs. 5-6 for distributions).

Annotation Protocol

Two annotators worked simultaneously in a shared spreadsheet to label each title with:

- Topic: one of five IPOLEKSOSBUDHANKAM categories (Ideology, Politics, Economy, Socio-culture, Security-Defense)
- Threat: binary label (threat / non-threat)

For the threat and news categorization label, annotators followed the operational definition provided in the Indonesian Defense White Paper (Ministry of Defense, 2015). A news title was labeled as a threat if it described or implied actions that could undermine national security, provoke violence, incite intimidation, or destabilize public order, in line with the IPOLEKSOSBUDHANKAM framework. Examples include reports of organized violence, explicit calls for attacks, or warnings of destabilizing actions. Conversely, figurative or non-security "threats" (e.g., "Minister threatens to resign") were excluded.

As annotation was conducted collaboratively in real time, inter-annotator agreement metrics such as Cohen's κ were not computed. Instead, disagreements were resolved by consensus during annotation sessions, ensuring consistency across labels. While this approach maintains efficiency and consistency, it limits the ability to quantify label reliability. Future studies will address this by employing independent annotators and reporting κ/α values.

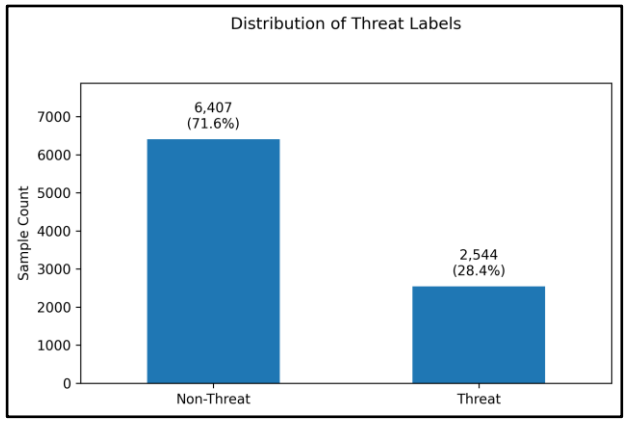

Fig. 5: Dataset distribution for threat categorization

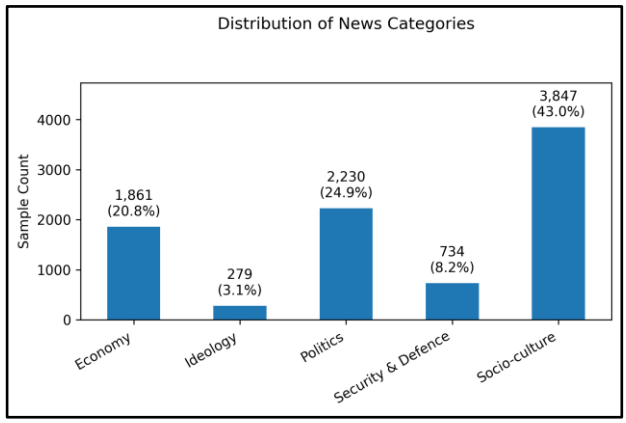

Fig. 6: Dataset distribution for news topic categorization

Class Distribution and Imbalance

The dataset exhibits a clear imbalance across both tasks. In the threat detection task, 6,407 titles were labeled as non-threat, while only 2,544 were identified as threat (Figure 5). A similar long-tailed distribution appears in the topic classification task, where Socio culture news dominates (3,847 articles), followed by Politics (2,230) and Economy (1,861). Security-

Defense accounts for 734 articles, while Ideology remains the most under-represented category with only 279 examples (Figure 6). These disparities highlight the challenge of training models that effectively capture patterns in minority classes.

No reweighting, resampling, or other balancing techniques were applied; models were trained directly on the naturally skewed distributions. This design was intended to evaluate model robustness under realistic class distributions. Nonetheless, imbalance mitigation methods such as focal loss, weighted objectives, or oversampling remain important avenues for future research.

Train/Validation/Test Split

We normalized titles (lowercasing, punctuation/whitespace stripping) and removed duplicate titles before splitting. The final split was 80/10/10 for train/validation/test. Because the collection window spans one week, we preserved temporal ordering and recorded outlet sources. To ensure robustness, all experiments were repeated with three random seeds (42, 123, 2025), and results are reported as mean ± standard deviation across these runs.

Data Pre-Processing

Before ingestion into Indo BERT, all titles underwent a minimal cleaning routine.

Embedded hyperlinks were stripped using regular expressions, and all Unicode emoji or dingbat characters were removed to ensure compatibility with the tokenizer.

The resulting plain-text strings were processed with the Indo BERT-base-uncased Word Piece tokenizer, which automatically lowercases text, appends the special tokens [CLS] and [SEP], and truncates sequences exceeding 36 sub-words. This threshold was selected because it covers approximately 95% of news titles, thereby preserving efficiency while minimizing information loss. No additional sampling, reweighting, or balancing techniques were applied; models were trained directly on the naturally imbalanced distributions. While this design ensured that evaluation reflects real-world skew, future work will incorporate strategies such as focal loss, weighted objectives, or oversampling to further improve minority-class performance.

Model Variants and Hyper-Parameters

Tables 2-4 summarizes the hyper-parameter configurations used across all model variants. For CNN-based hybrids, a single Conv1D layer was employed, with the number of filters varied among {300, 250, 200} during tuning, with kernel size fixed at 3. The configuration with 250 filters, shown in Figure 2, was selected as it provided the best balance of performance and efficiency in preliminary experiments. This choice is consistent with prior findings that mid-range filter sizes capture n-gram patterns effectively without excessive computational cost (Murfi et al., 2024).

Table 2: Hyper-parameters for each model variant

Model variant	Key layers / units	L2 kernel regularization
BERT-Threat	Dense = 128	0.001
BERT-Category	Dense = 128	0.001
BERT-MT	Shared dense = 256 Threat head = 128 Category head = 128	0.001
BERT-CNN	Conv filters = 300 / 250 / 200	0.001
BERT-LSTM	LSTM units = 300 / 200	0.001
BERT-Bi-LSTM	Bi-LSTM units = 300	0.001

Table 3: Per-class F1 of BERT (single-task) on topic classification (best run).

Model	Ideology	Politics	Economy	Socio-culture	Security-Defense
BERT-Category	0.65	0.89	0.84	0.87	0.95

Table 4: Per-class F1 of BERT (single-task) on threat detection (best run)

Model	Non-threat	Threat
BERT-Threat	0.92	0.78

During tuning, we evaluated 200 and 300 hidden units for LSTM and selected 300 for the final runs; Bi LSTM was fixed at 300 units per direction based on prior findings (Jiang et al., 2022). In the bidirectional variant, concatenation of forward and backward outputs yielded 600-dimensional representations.

Both single task and multi task heads use 128 units. Multi-task adds a shared 256-unit layer to encourage feature sharing while keeping task heads lightweight and comparable across CNN/LSTM/Bi-LSTM variants. This was a deliberate design choice to reduce overfitting and to ensure fair comparison across CNN, LSTM, and Bi LSTM hybrids (Guna Mandhasiya et al., 2024).

Optimization employed the Adam W optimizer with a learning rate of 3e 5, a batch size of 32, and an early stopping patience of 2 epochs, consistent with best practices reported in Transformer fine-tuning studies

(Devlin et al., 2019; Mosbach et al., 2021). Dropout (0.5) was applied following common strategies for stabilizing training in neural text classifiers.

Results

Single-task BERT shows clear variation across categories. The model performs best on Security-Defense (F1 = 0.95) and Politics (F1 = 0.89), while Socio-culture and Economy also perform strongly with F1 scores of 0.87 and 0.84, respectively. Ideology is the most difficult class, with F1 dropping to 0.65, reflecting its status as the most under-represented category in the dataset. In the best run, macro-F1 reached 0.84 across the five classes. The weighted F1 averaged across seeds is 0.84±0.01 (Table 5), which is a strong outcome given the imbalanced class distribution.

The model achieves high performance on the majority non-threat class (F1 = 0.92). However, performance on the minority threat class is notably lower (F1 = 0.78), underscoring the challenge of detecting rare but critical threat instances. This imbalance results in a weighted F1 of 0.87 across seeds (see Table 5).

As shown in Table 5, aggregated results across three random seeds confirm the stability of model performance. The single-task baselines achieve solid results, with BERT Category reaching 0.84±0.01 F1 and BERT-Threat 0.87±0.00 F1. Incorporating multi-task learning improves threat detection, with the best hybrid reaching 0.88, while others remain at 0.86 (comparable to the 0.87 achieved by the single-task model). Among the hybrids, BERT CNN achieved the highest threat F1 (0.88±0.01) while also maintaining strong category classification (0.86±0.01). BERT LSTM and BERT Bi LSTM showed comparable performance, with category F1 at 0.85 and threat F1 at 0.86. Overall, these results validate that multi-task training provides consistent gains on the minority threat class, while category classification remains stable across all architectures.

Table 6 reports training times as mean ± standard deviation across three seeds. Inference latencies are mean values. Among the multi-task variants, BERT-LSTM trained the fastest (153±2 s), followed by BERT-Bi-LSTM (161±24 s) and BERT-CNN (180±24 s). The plain BERT-MT required the longest time (193±22 s). Single-task baselines trained slightly faster overall: BERT-Category averaged 138±12 s and BERT-Threat 167±75 s. The larger variance for BERT-Threat reflects different early-stopping epochs across seeds.

Table 5: Weighted F1 scores of BERT-based models (mean ± std across 3 seeds) on news categorization and threat detection

Model	Cat Acc	Cat F1	Thr Acc	Thr F1
BERT-Category	0.84±0.01	0.84±0.01	-	-
BERT-Threat	-	-	0.87±0.00	0.87±0.00
BERT-MT	0.85±0.00	0.85±0.00	0.85±0.01	0.85±0.01
BERT-CNN	0.85±0.01	0.86±0.01	0.88±0.00	0.88±0.01
BERT-LSTM	0.86±0.00	0.85±0.00	0.86±0.00	0.86±0.00
BERT-Bi-LSTM	0.86±0.00	0.85±0.00	0.86±0.00	0.86±0.00

Table 6: Training and inference times of BERT-based models

Model	Training Time (s)	Inference (ms/item, b32)	Inference (ms/item, b1)
BERT-Category	138±12	2.13	66.72
BERT-Threat	167±75	2.14	68.20
BERT-MT	193±22	2.14	66.72
BERT-CNN	180±24	2.23	68.05
BERT-LSTM	153±2	2.16	68.05
BERT-Bi-LSTM	161±24	2.19	67.63

We report both throughput (batch size 32) and single-item latency (batch size 1). At batch size 32, all models process 450–470 items per second (2.1 2.2 ms per title). At batch size 1, latency is 67–69 ms per title, confirming near real-time feasibility. Differences across architectures are small: the CNN variant adds ~0.1 ms relative to the plain multi-task baseline, while LSTM and Bi-LSTM are 2.16–2.19 ms at batch size 32.

Together, these results confirm that the multi-task hybrids improve F1 scores (Table 5) while maintaining operational efficiency. Note that we did not apply imbalance-mitigation strategies (e.g., class weights or focal loss) and did not conduct statistical significance testing in this study (see Limitations). Inference latency was measured on an NVIDIA L4 (24 GB VRAM) on a host with 94 GB RAM using Tensor Flow (FP32), sequence length 36 sub words, and batch size 32.

Discussion

The experimental findings provide several insights into the effectiveness and practicality of BERT-based single-task and multi-task architectures for Indonesian news classification and threat detection. First, the single-task baselines established strong performance, with BERT-Category achieving a weighted F1 of 0.84±0.01

across five categories and BERT-Threat attaining 0.87±0.00 on binary threat detection. A closer per-class analysis revealed that Security-Defense and Politics were consistently the easiest categories to predict, while Ideology remained the most difficult due to its underrepresentation in the dataset.

Second, multi-task learning yielded consistent gains, particularly for the minority threat class. The BERT-CNN hybrid achieved the strongest overall performance (Threat F1 = 0.88±0.01), demonstrating the utility of convolutional layers in extracting local features relevant for detecting security-related signals. Both BERT-LSTM and BERT-Bi-LSTM improved threat F1 scores to 0.86, confirming that recurrent architectures can complement BERT embeddings, although their advantages were smaller and more variable. These results suggest that convolutional hybrids strike the best balance between accuracy, efficiency, and stability.

Third, the low variance across seeds (< 0.02 F1) underscores the robustness of the observed improvements, supporting the reproducibility of the comparative evaluation.

Finally, the efficiency analysis confirmed that the proposed architectures are feasible for deployment in real-world monitoring systems. All multi-task hybrids achieved 450 470 items per second at batch size 32 (2.1 2.2 ms per title), with single-item latency under 70 ms. these results indicate that multi-task models provide measurable improvements in threat detection without sacrificing inference speed. With training times under four minutes, the architectures can be effectively retrained and integrated into large-scale Indonesian news monitoring pipelines.

Limitations and Future Work

Despite careful design, several limitations remain. First, evaluation reliability was constrained. Although models were trained across multiple random seeds and reported as mean ± standard deviation, no bootstrap significance testing was performed, and inter-annotator agreement metrics such as Cohen's κ and Krippendorff's α were not computed. As a result, label reliability, particularly for the subjective threat category, cannot be fully quantified. Future work will incorporate bootstrap-based confidence intervals and independent annotation with agreement reporting to strengthen robustness.

Second, model comparisons were limited in scope. No dedicated ablation experiments such as freezing the encoder, varying token length, or removing hybrid heads were conducted, and only BERT based variants were considered. Broader baselines such as fastText, CNN Text, or lightweight LLMs would provide valuable context, while ablations would clarify the role of each design choice. These remain important directions for further investigation.

Finally, deployment factors were only partially evaluated. While class imbalance was documented and latency measured under both batch size = 32 and batch size = 1, no imbalance-mitigation strategies such as focal loss, oversampling or class-weighted objectives were applied, and memory footprint was not measured. Addressing these aspects in future work will strengthen robustness and operational feasibility on resource-constrained platforms.

Conclusion

This study presents a systematic comparison of BERT-based single-task and multi-task architectures for Indonesian news categorization and threat detection. The single-task baselines demonstrated strong performance, achieving weighted F1 scores of 0.84±0.01 for topic classification and 0.87±0.00 for threat detection. Multi-task learning consistently improved performance on the minority threat class, with the CNN-based hybrid obtaining the best results (Threat F1 = 0.88±0.01), while the LSTM- and Bi-LSTM-based hybrids achieved comparable improvements of around 0.86 F1. The low variability across random seeds (<0.02) further supports the robustness of these findings. Per-class analysis also indicates that Ideology remains the most challenging category (F1 = 0.65), consistent with its limited representation (279 samples).

In terms of efficiency, all evaluated models demonstrated operational feasibility. Training times ranged from 138±12 s (BERT-Category) to 193±22 s (BERT MT), with the LSTM variant converging the fastest. Inference benchmarks confirmed near real-time viability, with batch size 32 throughput of 450 470 items per second (2.1 2.2 ms per title) and single-item latency of 67 69 ms. These results demonstrate that multi-task BERT hybrids improve threat detection while maintaining inference efficiency, making them suitable for integration into large-scale monitoring systems for Indonesian online news.

Acknowledgment

We extend our gratitude to Bina Nusantara University for their support and encouragement, which played a vital role in fostering collaboration and enhancing the depth of our research.

Funding Information

This research was made possible through the financial support of Bina Nusantara University. The authors extend their sincere gratitude for the funding provided via the University's Research Fund, which substantially contributed to the execution of this study and supported the dissemination of its outcomes to the wider academic community.

Authors Contributions

A. Mustain Billah: Conducted the research, performed the analysis, and drafted the manuscript.

Sani M Isa: Supervised the research and reviewed the manuscript.

Ethics

The submitted manuscript constitutes original work by its authors and has not been published or is under consideration elsewhere. Each author has reviewed and approved its content, confirming its accuracy and adherence to academic guidelines. The study and its dissemination were carried out in strict adherence to ethical principles, with no conflicts of interest or ethical concerns arising at any stage. Moreover, the research fully conformed to the ethical guidelines prescribed by Bina Nusantara University, underscoring our commitment to responsible research.

Data Availability Statement

The dataset employed in this research was derived from Indonesian news RSS feeds that are publicly accessible. To promote transparency and reproducibility, both the list of RSS sources and the processed dataset have been made available at: https://github.com/039710/IPOLEKSOSBUDHANKAM.

References

Abas, A., Elhenawy, I., Zidan, M., & Othman, M. (2022). BERT-CNN: A Deep Learning Model for Detecting Emotions from Text. *Computers, Materials and Continua*, *71*(2), 2943–2961. https://doi.org/10.32604/cmc.2022.021671

Belaroussi, R., Noufe, S. C., Dupin, F., & Vandanjon, P.-O. (2025). Polarity of Yelp Reviews: A BERT–LSTM Comparative Study. *Big Data and Cognitive Computing*, *9*(5), 140. https://doi.org/10.3390/bdcc9050140

Crawshaw, M. (2020). Multi-Task Learning with Deep Neural Networks. *ArXiv*, 1–30. https://doi.org/10.48550/arXiv.2009.09796

Dai, W., Yu, T., Liu, Z., & Fung, P. (2020). Kungfupanda at SemEval-2020 Task 12: BERT-Based Multi-TaskLearning for Offensive Language Detection. *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, 2201–2206. https://doi.org/10.18653/v1/2020.semeval-1.272

Ministry of Defense. (2015). *White Paper*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv*, 1–14. https://doi.org/10.48550/arXiv.1810.04805

El Mahdaouy, A., El Mekki, A., Essefar, K., El Mamoun, N., Berrada, I., & Khoumsi, A. (2021, April). Deep multi-task model for sarcasm detection and sentiment analysis in Arabic language. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop* (pp. 334-339).

Faisal, D. R., & Mahendra, R. (2022). Two-Stage Classifier for COVID-19 Misinformation Detection Using BERT: a Study on Indonesian Tweets. *ArXiv*, 1–10. https://doi.org/10.48550/arXiv.2206.15359

Guna Mandhasiya, D., Murfi, H., & Bustamam, A. (2024). The hybrid of BERT and deep learning models for Indonesian sentiment analysis. *Indonesian Journal of Electrical Engineering and Computer Science*, *33*(1), 591–602. https://doi.org/10.11591/ijeecs.v33.i1.pp591-602

Hendrawan, R., Adiwijaya, & Al Faraby, S. (2020). Multilabel Classification of Hate Speech and Abusive Words on Indonesian Twitter Social Media. *2020 International Conference on Data Science and Its Applications (ICoDSA)*, 1–7. https://doi.org/10.1109/icodsa50139.2020.9212962

Jia, Q., Cui, J., Xiao, Y., Liu, C., Rashid, P., & Gehringer, E. F. (2021). ALL-IN-ONE: Multi-Task Learning BERT models for Evaluating Peer Assessments. *ArXiv*, 1–10.

Jiang, X., Song, C., Xu, Y., Li, Y., & Peng, Y. (2022). Research on sentiment classification for netizens based on the BERT-BiLSTM-TextCNN model. *PeerJ Computer Science*, *8*, e1005. https://doi.org/10.7717/peerj-cs.1005

Keya, A. J., Shajeeb, H. H., Rahman, Md. S., & Mridha, M. F. (2023). FakeStack: Hierarchical Tri-BERT-CNN-LSTM stacked model for effective fake news detection. *PLOS ONE*, *18*(12), e0294701. https://doi.org/10.1371/journal.pone.0294701

Lin, C.-H., & Nuha, U. (2023). Sentiment analysis of Indonesian datasets based on a hybrid deep-learning strategy. *Journal of Big Data*, *10*(1), 1–20. https://doi.org/10.1186/s40537-023-00782-9

Loshchilov, I., & Hutter, F. (2019). *Decoupled Weight Decay Regularization*. https://doi.org/10.48550/arXiv.1711.05101

Mehta, M., Gada, D., Sharma, R., Chavan, K., & Kanani, P. (2022). Offense Detection Using BERT and CNN. *2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT)*, 1–5. https://doi.org/10.1109/gcat55367.2022.9971862

Mosbach, M., Andriushchenko, M., & Klakow, D. (2021). *On the Stability of Fine-tuning BERT*. https://doi.org/10.48550/arXiv.2006.04884

Murfi, H., Syamsyuriani, Gowandi, T., Ardaneswari, G., & Nurrohmah, S. (2024). BERT-based combination of convolutional and recurrent neural network for indonesian sentiment analysis. *Applied Soft Computing*, *151*, 111112. https://doi.org/10.1016/j.asoc.2023.111112

Praha, T. C., Widodo, W., & Nugraheni, M. (2024). Indonesian Fake News Classification Using Transfer Learning in CNN and LSTM. *JOIV : International Journal on Informatics Visualization*, *8*(3), 1213. https://doi.org/10.62527/joiv.8.2.2126

Rai, N., Kumar, D., Kaushik, N., Raj, C., & Ali, A. (2022). Fake News Classification using transformer based enhanced LSTM and BERT. *International Journal of Cognitive Computing in Engineering*, *3*, 98–105. https://doi.org/10.1016/j.ijcce.2022.03.003

Ruder, S. (2017). *An Overview of Multi-Task Learning in Deep Neural Networks*. https://doi.org/10.48550/arXiv.1706.05098

Shah, P., Patel, H., & Swaminarayan, P. (2024). Multitask Sentiment Analysis and Topic Classification Using BERT. *ICST Transactions on Scalable Information Systems*, *11*, 1–12. https://doi.org/10.4108/eetsis.5287

Zhang, Y., & Yang, Q. (2021). A Survey on Multi-Task Learning. *ArXiv*, 1–30. https://doi.org/10.48550/arXiv.1707.08114